图书情报工作
第 65 卷 第 10 期 2021 年 5 月

社会科学数据的创建和使用研究*
——二次匹配数据采集规则的应用

陈欣[1] 曹朝金[2] 叶春森[1] 汪传雷[1]
[1] 安徽大学商学院 合肥 230009 [2] 合肥工业大学管理学院 合肥 230009

摘 要:[目的/意义]在数据生命周期框架下,创新性地提出一种从论文中采集社会科学数据创建和使用相关信息的方法,并深入研究其基本情况,为社会科学数据的研究提供新思路。[方法/过程]以学科交叉性较强的物流研究领域 2015 – 2020 年的 CSSCI 收录的论文为样本,通过迭代式方法构建基于数据生命周期的“泛化 – 精确关键词词库”,采集社会科学数据的相关信息,并结合社会科学数据外部环境信息,对社会科学数据的创建和使用进行全面研究。[结果/结论]在采集论文中社会科学数据的创建和使用相关信息上,二次匹配数据采集规则具有可行性和高效性,互联网已经成为社会科学研究主要的数据搜集方式,不同研究主题的数据使用偏好不同,对于数据分析工具的使用普及度仍然较低。

关键词:社会科学数据 泛化 – 精确词库 二次匹配数据采集规则 Python 文献计量
分类号:G203
DOI:10.13266/j.issn.0252 – 3116.2021.10.010

1 引言

数据被认为是促进各个领域创新的主要资源[1],尤其是在这个数据井喷式增长的时代,无论是商业领域,还是学术领域都在接受着大数据所带来的冲击。科学数据是科学研究过程的重要组成部分,既是科学研究的成果,也是科学研究的基础。学者们越来越重视数据驱动研究(Data-driven research),尤其是在生命科学、地球科学和地理科学等自然科学领域中[2]。为了使科学数据的管理更加规范成熟,国务院办公厅于 2018 年 3 月 17 日印发了《科学数据管理办法》(以下简称《办法》)[3],但《办法》主要面向自然科学、工程技术科学等领域,暂未对社会科学(以下简称“社科”)领域研究中的科学数据管理作出明确规定。

社科领域的学者们近年来不断使用大规模数据分析方法、复杂的数学模型和丰富的数据分析工具等[4],科学数据在该领域也有着重要的价值和作用,社会科学数据的管理同样需要得到重视。社会科学数据可以是广义的与社科领域有关的数据,例如社科调查数据、政府统计数据、商业公开数据等,也可以指狭义的社科领域研究活动所产生的各种数据,例如文本记录、数值统计、图像数据等[5]。正是由于社会科学的研究方法和数据格式多种多样,并非都是数值型数据,还包括文本数据、档案数据、汇编数据、PDF 格式数据等,还包括微观尺度数据和宏观尺度数据,缺乏统一标准,导致该领域中科学数据的利用状况较差,并且分散在各个研究者和组织的手中,社会科学研究体现出多样性和不确定性[6],加之目前对社会科学数据的研究较少,对社会科学数据的特征了解不够深入,在出台制订社会科学数据管理政策上缺少现实依据,阻碍了社会科学数据的管理与服务。本研究以社会科学研究中物流研究领域为落脚点,以 CSSCI 收录的论文作为样本,在数据分析方面,首先从文献计量的角度分析社会科学数据的外部环境,同时基于数据生命周期框架,利用二次匹配规则采集论文中的社会科学数据相关信息,并具体分析其创建和使用的特点,结合外部环境深入讨论不同发文单位、不同研究热点主题对社会科学数据的使用偏好关系,分析社会科学数据的创建和使用规律。

* 本文系国家社会科学基金青年项目“学术大数据环境下社会科学数据开放的影响因素及评价研究”(项目编号:19CTQ029)和安徽高校人文社会科学研究重点项目“安徽省物流科技数据使用现状与对策研究”(项目编号:SK2017A0016)研究成果之一。
作者简介:陈欣(ORCID:0000 – 0001 – 7528 – 0789),讲师,博士,E-mail:chenxinnju@foxmail.com;曹朝金(ORCID:0000 – 0003 – 2683 – 2051),硕士研究生;叶春森(ORCID:0000 – 0001 – 7782 – 2721),副教授,博士;汪传雷(ORCID:0000 – 0003 – 4498 – 3152),教授,博士。
收稿日期:2020 – 09 – 24 修回日期:2021 – 02 – 23 本文起止页码:90 – 104 本文责任编辑:王传清

2 社会科学数据研究现状

2.1 社会科学数据的管理和服务研究

当前社会科学数据的管理和服务研究较少，相比较而言，具有明显特征的自然科学数据的相关研究较多，因此无论是国内还是国外都有指导其管理的政策，如我国的《办法》[3]、美国国家航空航天局(National Aeronautics and Space Administration，NASA)的 *Data & Information Policy*[7]、英国生物技术与生物科学研究理事会(Biotechnology and Biological Sciences Research Council，BBSRC)的 *BBSRC Data Sharing Policy*[8] 等。也正因如此，我国开放科学数据共享的资源类型较为单一，以自然科学数据为主[9]。目前我国在社会科学数据管理的认知与做法上存在一些问题，例如将社会科学数据管理简单化地等同于常规性的资料工作[5]。因此对社会科学数据创建和使用的研究显得极为必要，只有在充分了解社会科学数据创建和使用特点的基础上，才能制定出面向社会科学研究的科学数据管理和服务政策。

2.2 社会科学数据特征和性质研究

关于社会科学数据特征和性质的研究，国内外学者多以某个数据库或元数据仓库(Data Citation Index，简称 DCI)为样本，如罗鹏程等基于 DataCite 分析科学数据在时间、空间等维度上的特征[10]；孟祥保等分析 DCI 中历史学、教育学等 5 个学科的科学数据的结构特征[11]。也有部分学者期刊论文为样本，以人工方式对文本进行内容分析进而收集论文中的科学数据相关信息，如沈婷婷以《中国社会科学》发表的论文为样本，统计分析了研究者获取科学数据的途径、科学数据的类型等，并对图书馆的科学数据服务提出了建议[12]。综上所述，现有研究仍存在一些不足：一方面，社会科学研究对于如 DataCite、DCI 等科学数据库使用频率较少，提交共享科学数据则更加稀少。大多数社会科学研究人员的数据存储地点为个人计算机，他们的分享数据的方式主要通过非正式渠道进行，而使用存储库共享数据的只占 46%[13]。因此以科学数据库为样本分析我国社会科学数据的特征可能缺乏一定代表性；另一方面，论文作为科学研究结果的表现形式，从中可以提炼出社会科学数据相关信息，但是目前以论文为样本进行的研究，多以人工的方法对论文文本进行内容分析，一定程度上限制了样本的数量及信息提取的精确度，在研究方法上存在一定局限性。

2.3 Python 在文本分析方面的应用研究

Python 作为当下最热门的一种计算机编程语言，在科学研究中也常用于文本分析，如谭春林等通过 Python 编程对期刊论文的文本内容进行挖掘[14]；张娜等采用 Python 中的 snowNLP 模块对文本数据进行意见挖掘，将文本数据归为积极和消极两类[15]；刘玉林等利用 Python 对电商的在线评论进行文本情感分析[16]。从上述研究中可以看出，Python 的文本分析能力已经日趋完善，因此基于 Python 的特点和应用情况，本研究利用 Python 编写程序对论文进行内容分析，从样本论文中采集出社会科学数据的相关信息。

基于现有研究的不足，同时考虑到 Python 在文本分析方面的广泛应用，本研究提出一种采集社会科学数据相关信息的逻辑思维：首先通过迭代式方法构建一种基于数据生命周期的“泛化-精确关键词词库”，进而依此设计一种基于 Python 的二次匹配数据采集规则，该规则结合词库可以从论文中高效采集社会科学数据的相关信息。

3 研究设计

3.1 研究思路

研究思路如图 1 所示，本研究首先从 CSSCI 中检索出指定区间段内的物流领域文献作为研究样本，在基于数据生命周期的数据采集框架下，构建词库，通过词库和二次匹配数据采集规则获取社会科学的创建和使用特点数据，并利用文献计量分析方法对社会科学数据所处的外部环境进行分析，包括发文单位、发文作者、发文时间、研究热点。然后利用统计分析方法对社会科学数据的创建和使用特点进行分析，具体包括创建主体、创建方法、数据类型、数据分析方法、数据分析工具、模糊词的使用 6 个方面。在分析社会科学数据的创建和使用特点时，结合社会科学数据的外部环境分析角度，更加全面地研究社会科学数据在创建和使用维度上的特点。

3.2 样本来源

本研究选择 CSSCI 来源期刊收录的物流研究领域论文作为样本，主要基于以下几点原因：①论文作为科学研究结果的表现形式，一般包含对科学数据从搜集到使用全过程的描述，便于提炼与科学数据创建和使用相关的信息；②在学术领域具有代表性的 CSSCI 经常是学者们的首选样本[17-19]；③本研究选取社会科学研究领域中，具有跨学科、跨行业特性的物流研究论文

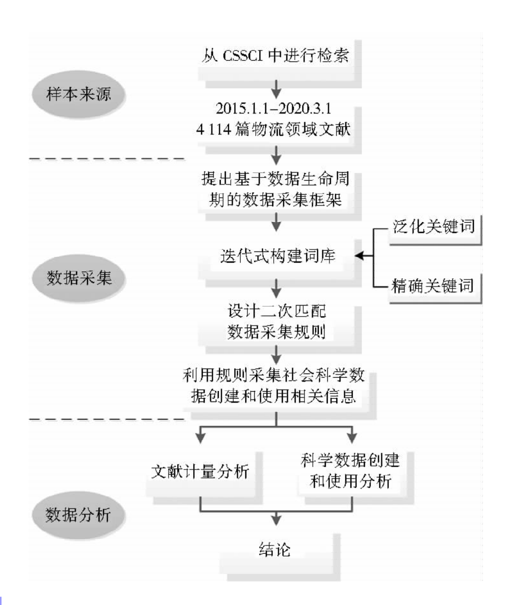


图 1 研究思路

为样本，对社会科学研究有一定代表性，以物流研究领域为试点，为后续研究社会科学中其他学科的科学数据创建和使用奠定基础。

为了提高查全率和查准率，本研究从中国知网(CNKI)、万方、重庆维普数据库中，以物流 7 个功能要素"运输、仓储、装卸搬运、包装、流通加工、配送、信息处理"加上"物流"和"供应链"进行主题检索，其他检索条件设置如表 1 所示，剔除重复与无关的论文后，共得到样本论文 4 114 篇。

表 1 检索条件

名称	限制条件
文献类型	期刊
主题	物流 7 个功能要素、物流、供应链
地理范围	不限
检索时段	2015 年 1 月 1 日 – 2020 年 3 月 1 日
来源期刊	CSSCI 收录期刊

3.3 数据采集

本研究构建一种基于数据生命周期的"泛化 – 精确关键词词库"、一种迭代式构建词库的方法，同时提出一种基于 Python 的二次匹配数据采集规则，规则利用词库中的两类关键词对论文内容进行检索实现二次匹配过程，进而采集社会科学数据的相关信息。

3.3.1 基于数据生命周期的"泛化 – 精确关键词词库"

数据生命周期是一个周而复始、动态变化的过程，一般来看，数据会经历从出现到被使用到最后消逝等几个环节[20]。目前国内外对数据生命周期理论的研究都较为完善，包括针对数据生命周期理论本身的研究，以及基于数据生命周期研究科学数据的特征和管理服务，例如国内的学者孟祥保[11]、丁宁[21]、武彤[22]等，国外的弗吉尼亚大学[23]、加利福尼亚大学[24]、新墨西哥大学[25]、英国数字管理中心(DCC)[26]、英国社会科学数据存储(UKDA)[27]等都针对不同学科或领域提出了相关的数据生命周期模型。本研究参考已有数据生命周期模型，最终确定从数据生命周期中的数据创建和数据使用两个维度构建采集框架，数据创建是指社会科学数据的产生与搜集环节，其具体内容包括创建主体、创建方法和数据类型；数据使用是指研究中对社会科学数据的具体分析环节，其具体内容包括数据分析方法、数据分析工具和模糊词的使用。基于数据生命周期的社会科学数据创建和使用信息的采集框架如表 2 所示，在此框架下，依据不同的数据采集单元构建词库，匹配采集文献中社会科学数据创建和使用相关信息。

词库包含"泛化"和"精确"两类关键词，"泛化关键词"指期刊论文中描述研究样本、资料、数据时的常用词，而这类词的使用场景又不仅局限于此，因此以"泛化关键词"进行检索的主要目的是缩小文本检索范围；"精确关键词"指在"泛化关键词"检索的基础上进一步精准定位社会科学数据相关信息的关键词。例如为了判断一篇论文的社会科学数据是否来源于统计局时，所用的词库中"泛化关键词"为"统计局"，但是文中出现"统计局"并不意味着这篇论文一定从统计局中获取数据。通过阅读文献，发现当上下文包含"统计局"，同时含有"数据来源""获取""查阅"等"精确关键词"时，则可确认该论文的数据来源为统计局。

3.3.2 迭代式构建词库

采集不同研究领域、不同采集单元的科学数据相关信息需要不同的词库，为了保证词库的完整性，本研究提出了一种迭代式构建词库的方法，见图 2。

(1)建立初始词库。基于前期对安徽省高校近 10 年来在 CSSCI 上发表的物流研究文献中的科学数据相关信息进行内容分析，并结合各采集单元的前期研究，初步确定了泛化关键词，利用 Python 提取样本论文中含泛化关键词的语句，并对语句进行分词处理，对与泛化关键词共现频数较高的词语进行判断比较，初步确定精确关键词，将泛化关键词和精确关键词存入词库中，形成初始词库。

表 2　基于数据生命周期的“泛化－精确关键词词库”

数据采集维度	数据采集单元	含义	泛化关键词	精确关键词	泛化关键词总计/个	精确关键词总计/个
数据创建	创建主体	谁创建该数据，如个人、政府机构、研究团队、专业调查公司、企业等	统计局 …	数据来源 获取 …	30	79
	创建方法	采集数据所用方法，如网络查找、实验（计算机模拟）、访谈、问卷、文献等	年鉴 …	本文 根据 …	35	75
	数据类型	按数据获取方式划分为一手数据和二手数据	问卷 …	本研究采用 收集数据 …	33	80
数据使用	数据分析方法	分析处理数据所使用的方法，如统计分析、数学模型	案例分析 …	采取 运用 …	21	16
	数据分析工具	各种数据处理软件等	Matlab …	使用 利用 …	69	10
	模糊词的使用	统计“绝大多数”“差不多”“大量”等8类模糊词的出现次数	绝大多数 …	– –	8	0

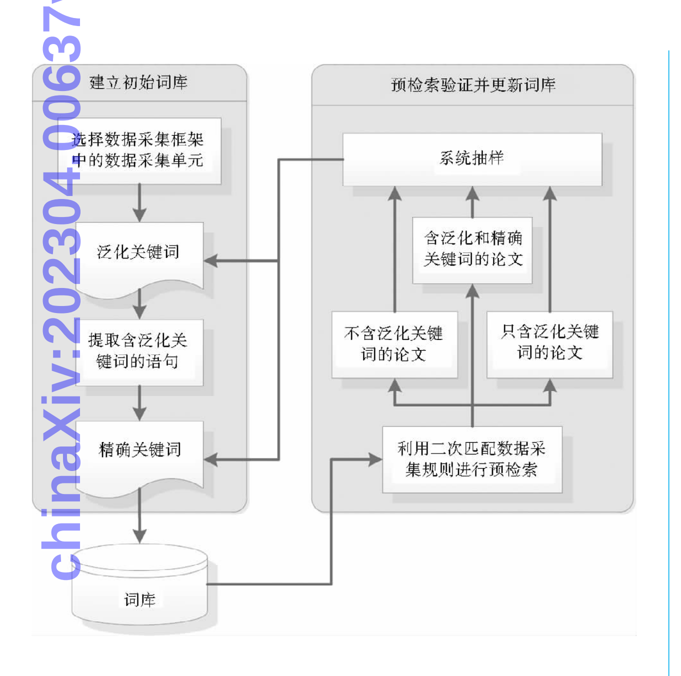


图 2　泛化－精确关键词词库建立流程

（2）预检索验证并更新词库。依据初始词库利用二次匹配数据采集规则进行预检索，得到 3 类论文：①不含泛化关键词的论文；②只含泛化关键词的论文；③含泛化和精确关键词的论文。对 3 类论文分别进行系统抽样，抽取 20% 的论文。对于①类论文，利用文本内容分析，判断并补充新的泛化关键词；对于②类论文，通过分析其上下文补充精确关键词；对于③类论文，分析其上下文判断精确关键词是否有效，若精确关键词匹配准确率低于 90%，则删除该精确关键词。若抽取的 3 类论文中没有新的泛化关键词以及精确关键词可以添加，同时所使用的精确关键词都有效，则跳出迭代，输出该类数据采集单元的词库。

最终得到基于数据生命周期的泛化－精确关键词词库，见表 2。

3.3.3　基于 Python 的二次匹配数据采集规则

针对科学数据的二次匹配数据采集规则，规则逻辑如图 3 所示：

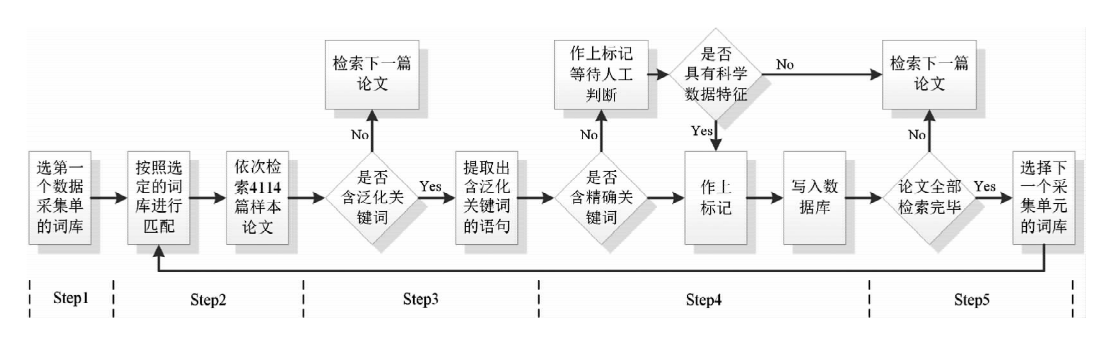


图 3　二次匹配数据采集规则

步骤 1:选择第一个数据采集单元的词库。

步骤 2:以词库中的“泛化关键词”依次对 4 114 篇论文的全部内容进行第一次检索。

步骤 3:判断论文是否含有“泛化关键词”,若是,则提取包含“泛化关键词”的语句,录入 Excel;反之检索下一篇论文。

步骤 4:以“精确关键词”进行检索,判断是否含有,若是,给该篇论文作上标记,具有词库所假定的科学数据信息,并将标记信息写入数据库中;反之,作上等待人工判断标记,由人工判断是否具有词库所假定的科学数据信息,若是则同样作上标记并写入数据库;反之,检索下一篇论文。

步骤 5:判断 4 114 篇论文是否全部检索完毕,若是,选择下一个数据采集单元的词库,返回步骤 2;反之,检索一下篇论文。

4 社会科学数据外部环境分析

分析社会科学数据的外部环境,即对社会科学研究领域文献的发文单位、发文作者、发文时间、关键词突显、研究热点进行分析。分析其外部环境,一方面有助于了解物流研究的基本情况,另一方面,结合外部环境分析,有助于深入了解社会科学数据的创建和使用。

4.1 发文单位分析

对发文第一单位进行分析发现,高等院校占比 97%,其他单位如公司、研究院、政府部门等占比 3%,说明社会科学研究的主体是高等院校。

进一步分析发文量在前 20 名的高校,如图 4 所示,北京交通大学发文排名第一,随后是重庆大学、上海海事大学、西南交通大学、中南大学、大连海事大学、中国人民大学、北京物资学院、东北大学、西安交通大学。这些高产院校的研究在一定程度上代表着国内物流领域的发展前沿,下文将结合高产院校在科学数据的使用与其科学内部特征进行分析。

4.2 发文作者分析

样本论文涉及作者共 10 138 人,平均每篇论文由 2.47 人完成。具体分布如表 3 所示,80% 的论文由 2 –4 人完成,说明物流领域的学术研究倾向于合作完成,其中 2 –3 人合作情况居多。

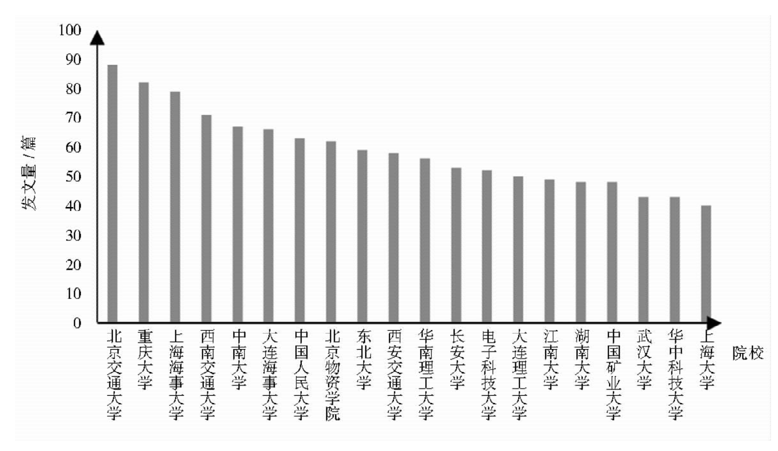


图 4 高等院校发文量排名(Top20)

表 3 论文合作人数占比

作者人数/人	论文数量/篇	所占比例/%
1	680	16.54
2	1 549	37.67
3	1 310	31.86
4	456	11.09
5	98	2.38
6	18	0.44
8	1	0.02

表 4 2015 年 1 月 –2020 年 3 月 CSSCI 高产作者

序号	第一作者	论文数量/篇	序号	第一作者	论文数量/篇
1	宋华	19	11	李健	11
2	唐建荣	16	12	黎继子	10
3	梁雯	14	13	冯颖	10
4	王道平	13	14	耶晓峰	9
5	但斌	12	15	张学龙	9
6	王文宾	11	16	王静	9
7	张建军	11	17	汪传雷	9
8	浦徐进	11	18	康凯	8
9	李新然	11	19	葛显龙	8
10	颜波	11	20	于辉	8

通过对高产作者进一步分析,如表 4 所示,发文量最多的是宋华(中国人民大学),重点研究供应链金融;排名第二的是唐建荣(江南大学),研究内容偏向

于区域物流和绩效评价；梁雯（安徽大学）以农村物流和协调发展等为研究主题，发文量排名第三；物流领域的高产作者还有王道平、但斌、王文宾、张建军、浦徐进、李新然、颜波等。

4.3 发文时间分析

一个学科领域的发展速度和发展程度可以从文献的年代分布中看出来。因为检索截止时间为 2020 年 3 月 1 日，且院校单位发表的论文占总论文的 97%，因此重点讨论院校单位及其发表的论文数量，如图 5 所示：

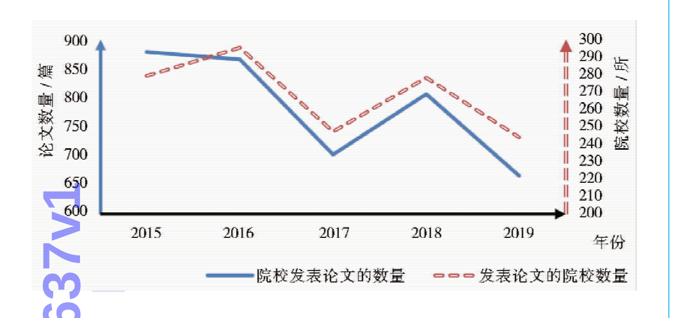


图 5 2015-2019 年院校发表论文数量及发表论文的院校数量

2015-2019 年之间论文数量的演变可分为 3 个阶段：第一阶段是 2015-2017 年，全国院校在物流领域中发表的论文以及在物流领域发表论文的院校数量总体都呈下降趋势；第二阶段为 2017-2018 年，受《商贸物流发展"十三五"规划》《关于积极推进供应链创新与应用的指导意见》《商务部等 8 部门关于开展供应链创新与应用试点的通知》等重大政策出台，供应链创新、物流降本增效等研究热门主题，推动了物流研究领域的发文量以及院校的关注度；第三阶段为 2018-2019 年期间，两者数量均有回落，一方面因为高水平期刊控制发文数量，另一方面由于前期物流热点研究密集，后期发文量呈现震荡下降。总体来看，行业政策出台、C 刊发文数量收紧、研究成果向国际期刊转移等因素均在一定程度上导致物流研究领域发文数量下降。

4.4 关键词突现性分析

通过研究关键词在不同时期的兴衰，可以用于探索一个研究领域内过去的潮流和未来的趋势，便于探究 2015-2019 年论文数量波动的原因。本研究利用 CiteSpace 软件对关键词进行突现性检测（Burst detection），突现强度越强，说明该关键词在这段时期内受到的学术关注度越突出[28]，如表 5 所示。

受经济环境和政策影响，2015 年与物流业相关的多种研究主题集中出现，例如电子商务、食品安全、绿色物流等，受其影响当年发文量为近 5 年最高。自 2016 年起，研究主题向物流信息、物联网、"互联网 +"等主题集中，下降至 12 个，其中 8 个关键词的热度没有维持到 2017 年，进一步导致了 2017 年发表的论文数量降低。2018 年物流产业、物流服务供应链等研究对象热度上升，并且 2017 年碳交易、碳税等关键词热度延续到了 2018 年，使得 2018 年论文发表数量有所上升。2019 年具有高突现强度的关键词都是从 2017、2018 年中延续而来，当年并没有新增高突现强度的关键词，故 2019 年发文数量有所回落。

表 5 2015-2020 年物流领域关键词突现分析

关键词	数量	平均突现强度	起止时间	起止时间示意图
电子商务、食品安全、绿色物流、物流网络、经济增长、低碳物流、应急管理、食品供应链、农产品物流、流通业、绩效评价、供应链协同、突发事件、区域经济、协同发展、结构方程模型	16	3.0884	2015	
回购契约、第三方物流、随机需求、协同	4	3.0945	2015-2016	
激励机制、物流信息、期权契约、商业信用	4	2.7785	2016	
物联网、碳减排、"互联网 +"、碳交易	4	3.2625	2016-2017	
微分博弈、制造业、碳税、收益共享	4	3.5012	2017-2018	
定价决策、长江经济带、公平偏好	3	3.3579	2017-2020	
物流产业、物流服务供应链、风险规避、城市交通、定价策略	5	3.5871	2018-2020	

4.5 物流研究热点分析

关键词是对一篇文献主题和内容的高度概括，因此通过对一个研究领域中的文献进行关键词共现分析，可以总结得到该领域的研究热点[29-32]。利用 CiteSpace 软件自带的剪枝算法 Pathfinder、Pruning sliced networks、Pruning the merged network，对 2015-2019 年间物流研究出现的关键词进行共现分析。并形成可视化的知识图谱，探寻物流研究中的热点领域[33]。

如图 6、表 6 所示，结合高产作者以及他们的研究

方向，可以将近 5 年物流领域的研究热点概括为：①跨境物流；②闭环供应链；③供应链金融；④农产品物流；⑤绿色物流。

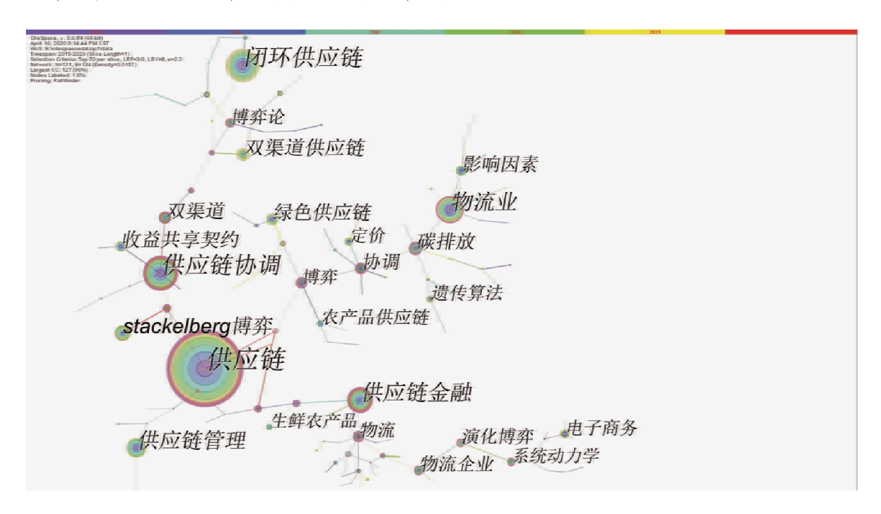


图 6　物流领域关键词共现知识图谱

表 6　物流领域高中心度关键词

关键词	中心度	频次	关键词	中心度	频次
博弈	0.9	41	供应链协调	0.69	157
信息共享	0.83	28	协调	0.69	53
区块链	0.83	12	碳排放	0.56	61
物流	0.82	44	双渠道	0.51	60
信息不对称	0.82	27	风险规避	0.48	21
供应链	0.81	330	定价决策	0.47	20
供应链金融	0.79	114	跨境电商	0.46	24
风险	0.76	6	物流企业	0.38	49
去中心化	0.75	2	演化博弈	0.33	44
资金约束	0.74	38	系统动力学	0.3	41

5　社会科学数据创建和使用分析

基于上文所提出的数据采集方法，获取论文中科学数据创建和使用的相关信息，并结合上文的外部环境分析，对其进行深入分析。

5.1　数据创建

5.1.1　科学数据来源分析

（1）科学数据来源情况概述。数据来源是指研究中科学数据的来源渠道，在样本分析基础上，将现有研究的数据源划分为科学调查、政府公开和商业公开 3 个来源渠道，见表 7。

表 7　数据来源类别、具体含义及使用情况

数据来源类别	具体特征	使用次数
科学调查数据	主要指通过社会科学研究方法由研究者自身或委托他人对研究对象开展调查获取的数据，该类数据多为一手数据，多以问卷、表格以及多种形式存在于研究者的电脑中	2 886
政府公开数据	来源于国家或地方政府发布的统计年鉴、统计公报、政策文件、报告等	1 560
商业公开数据	来源于行业协会、论坛会议发布的报告，企业公开、科学数据库的物流相关数据等	797

通过对 3 类数据源进行统计，使用次数由低到高依次为商业公开数据、政府公开数据、科学调查数据。

从数据获取的难易程度来看，政府公开数据和商业公开数据由于开放程度较高，数据格式较为一致，因此获取较易。科学调查数据因为其调查方法、调查对象、调查目的不同，获取难度较大，且数据格式多种多样，呈现出多、小、散的特点。尽管科学调查数据获取难度较大，但因其具有较强的自主性和研究内容针对性，也更受科研人员的青睐。

（2）高产院校的数据来源分析。高产院校的研究通常能体现一个研究领域的发展前沿，他们在科学数据使用上的特征具有一定代表性，根据上文对发文单位的分析，选取发文量在前 10 位的院校单位进行统计分析，如图 7 所示。每个高产院校都有一半以上的研究使用的科学数据的来源为科学调查数据，其中东北大学发表的论文有 95% 都使用了科学调查数据，科学调查是物流研究乃至社会科学研究领域中数据的主要来源方式。同时发现在高产院校中，北京交通大学、中国人民大学、北京物资学院这 3 所地处北京的高校，所使用的科学数据多来源于政府公开和商业公开，这一方面体现出北京的高校与政府之间的合作更为紧密，相比其他省份的高校更容易获取到政府数据，另一方面也说明了北京高校更加重视对政府公开数据的利用，对相关政策文件敏感度更高。

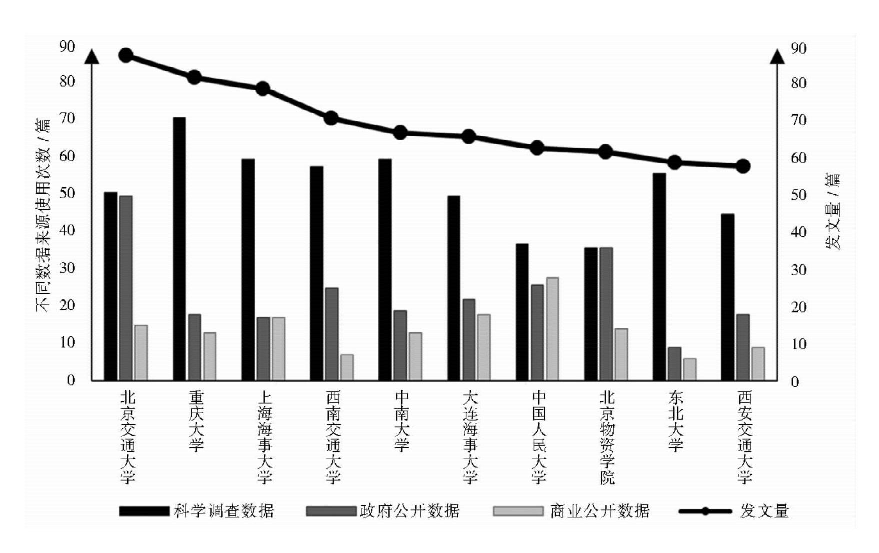


图7　发文量前10位的单位的数据来源情况

5.1.2　科学数据搜集方法分析

(1)科学数据搜集方法情况概述。数据搜集方法是指学者们通过何种途径搜集科学数据用于研究，本研究将科学数据的搜集方法分为网络查找和非网络查找两种。网络查找主要指通过互联网搜索专业数据库、行业报告、上市企业公开数据、统计年鉴、政府文件以及其他网站上数据等物流相关信息，非网络查找主要指通过仿真实验、发放问卷、开展调研、发起访谈的方式获取相关数据。本研究对样本文献的统计结果显示，使用网络查找搜集数据的论文占比56.52%，非网络查找占比50.58%，总体来看，网络搜集已经成为物流研究获取科学数据的主要途径。

(2)不同年份数据搜集方法选择分析。为了进一步了解研究者的数据搜集习惯的变化趋势，进一步分析不同年份网络搜集与非网络搜集的占比情况，如图8所示。通过分析发现，非网络搜集多年来在研究中的占比基本稳定，而网络查找搜集方式有不断升高的趋势，提高了近15%。一方面表明伴随计算机网络技术不断进步，已经有越来越多的研究人员利用网络来帮助自己完成科学研究过程中的数据搜集工作，数字化科研(E-Science)是社科领域科研人员的主要科研环境之一；另一方面，传统的社会科学调查等传统方法在现有研究中仍然占有重要地位。

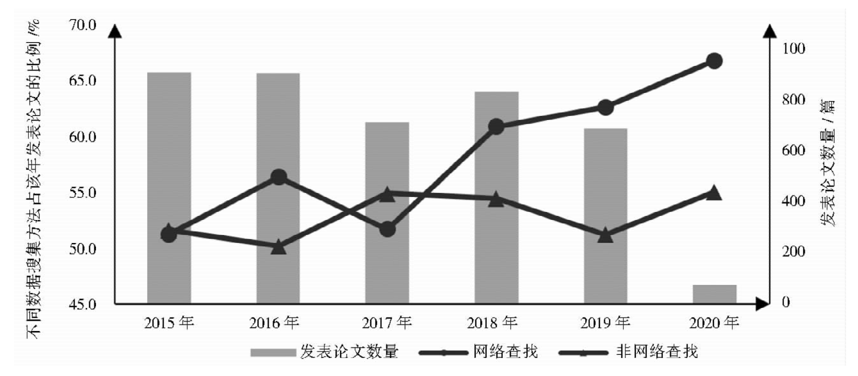


图8　不同数据搜集方法在各年份发表的论文中的占比

5.1.3　科学数据类型分析

(1)科学数据类型情况概述。目前对数据分类的研究较多，本研究依据数据产生的目的不同将科学数据划分为一手数据和二手数据，一手数据是指研究者通过实验、访谈、发放问卷等方式首次亲自收集并经过加工处理的数据，二手数据是指来源于他人调查和科学实验的数据[34]。另一方面，从数据格式看，可以将其分为文本、数值、图片、音频型等[11]，具体类别分布见表8。

总体来看，研究中二手数据的使用要略多于一手数据，前者占样本论文总数的75%，后者占65%，同时大部分数据类型的格式以文字和数值为主。一手数据中模型参数数据、算例数据和仿真数据占比较大，模型参数数据是在构建模型时所设定的前提条件数据；算

表 8 物流科学数据类别细分及在样本中的使用情况

数据类别	论文数量/篇	具体类别	数据格式	出现次数	数据类别	论文数量/篇	具体类别	数据格式	出现次数
一手数据	2 683	算例数据	数值	1 162	二手数据	3 077	专著数据	文本	1 724
		模型参数数据	文本、数值	1 035			政府文件数据	文本、数值	1 178
		仿真数据	数值	892			博硕论文数据	文本、数值	983
		问卷数据	文本、数值	439			统计年鉴数据	数值	631
							企业公开数据	数值	478
		调研数据	文本、数值、音频、图片	282			行业报告数据	文本、数值	383
							数据库数据	数值	349
		访谈数据	文本、数值、音频	242			统计局数据	数值	280
							统计公报数据	数值	134
		专家评价数据	文本、数值	225			其他网站数据	文本、数值、图片	107
		总计		4 277			总计		6 274
		样本论文数量/篇					4 114		

例数据指的是论文使用算例来验证文中提出的模型或相关结论的正确性时所使用到的数据，该类数据一部分为企业或其他部门的真实数据，另一类是由研究者根据模型条件设置的数据；仿真数据是在仿真实验中所应用的参数数据；除此之外常见的一手数据还有问卷数据、访谈数据等。另外，几乎每篇文献所进行的研究均会利用到期刊文本数据，故未在表 8 中列出。

(2)不同研究热点数据类型使用偏好分析。选取跨境物流、闭环供应链、供应链金融、农产品物流、绿色物流 5 个研究热点的相关论文，进一步探讨不同研究热点所使用的科学数据类型情况，如图 9 所示。研究跨境物流的论文中使用二手数据的次数要远远多于使用一手数据，而在闭环供应链的研究领域中更多的论文偏向于使用一手数据，供应链金融、农产品物流和绿色物流研究领域中一手数据和二手数据的使用次数差距较小。

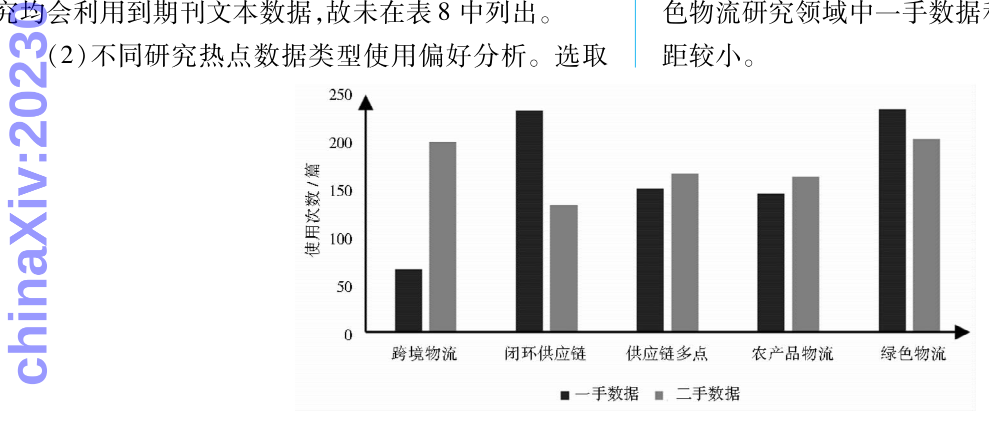


图 9 不同物流研究热点使用数据类型情况

从科学数据的具体类型来看，如图 10、图 11 所示，不同研究热点对一手科学数据的使用偏好明显不同，跨境物流偏好使用问卷数据；闭环供应链、绿色物流两个研究领域中，都倾向于使用模型参数数据、算例数据或仿真数据。从二手数据来看，跨境物流研究偏好政府文件数据，供应链金融偏好企业公开数据，农产品物流、绿色物流研究偏好于专著类数据。各研究热点间对于科学数据的使用偏好区别较大，究其原因，不同研究热点由于其研究基础、研究对象、研究方法不同，直接影响了科学数据的使用偏好，且该偏好伴随研究热点的深入而动态变化。

5.2 数据使用

5.2.1 科学数据分析方法分析

(1)科学数据分析方法情况概述。数据分析方法是指研究中进行数据处理和分析时所用方法，通过对样本中出现的数据分析方法进行统计，如表 9 所示，其中算例分析、实验法、统计学方法是使用最多的 3 种数据分析方法。说明在物流领域更多的是结合现实情况或者案例进行研究，而非单纯地理论研究，并且定量研究多于定性研究。大数据建模方法是指针对大数据分析所发展起来的一系列机器学习算法，例如决策树、支持向量机、人工神经网络等[35]。其他分析方法包括内容分析法、社会网络分析、文献计量分析等[36-37]。

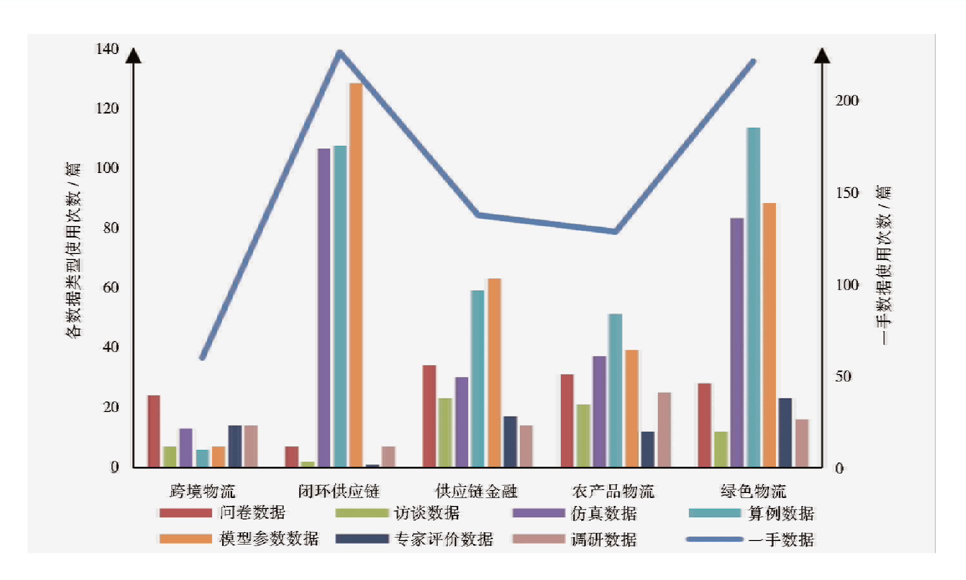


图 10　不同研究热点使用一手数据类型情况

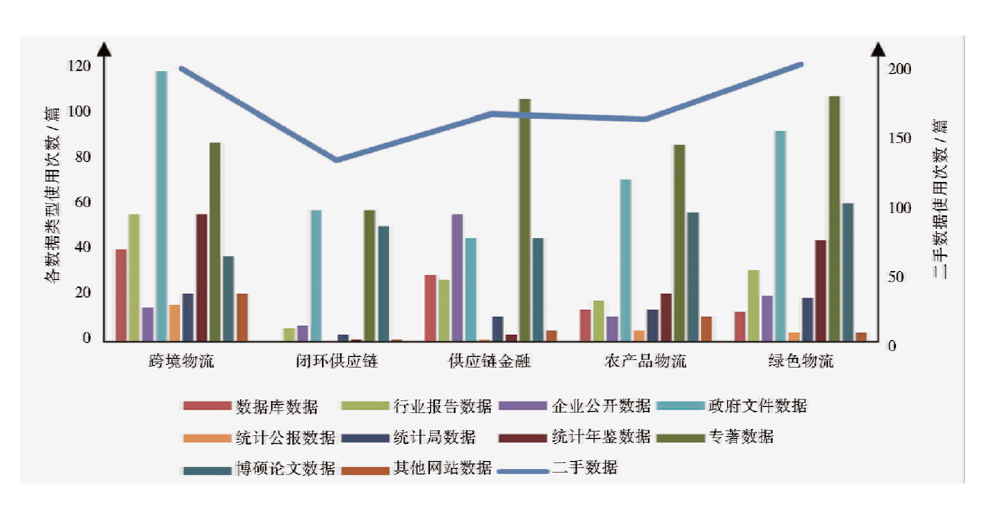


图 11　不同研究热点使用二手数据类型情况

表 9　数据分析方法类别及在样本中的使用情况

序号	数据分析方法	使用次数	序号	数据分析方法	使用次数	序号	数据分析方法	使用次数
1	算例分析	1 162	4	博弈分析	682	7	大数据建模方法	306
2	实验法	922	5	比较研究法	476	8	数据包络分析	158
3	统计学方法	911	6	案例研究法	349	9	其他分析方法	94

（2）不同物流研究热点数据分析方法的偏好分析。结合跨境物流、闭环供应链、供应链金融、农产品物流、绿色物流 5 个研究热点的论文数据分析方法的使用偏好分析，如图 12 所示，可以看出，统计学方法在供应链金融、跨境物流、农产品物流中都比较常用，但在闭环供应链中用得非常少，闭环供应链更偏好于算例分析、实验法和博弈分析。这也符合闭环供应链的研究特点，闭环供应链多侧重于理论研究，因此需要算例分析来验证理论结果，而其他研究热点则侧重于实证。

5.2.2　科学数据分析工具分析

（1）科学数据分析工具情况概述。分析工具是研究者进行数据分析选择的必然结果，分析工具可以协助、加快研究工作的开展，为研究工作带来极大的便利。对论文中提及的数据分析工具进行统计分析，如表 10 所示。其中 MATLAB、SPSS、AMOS 等仿真与统计分析软件使用次数较多。

（2）不同年份数据分析工具的选择分析。如表 11 所示，近 5 年 MATLAB 的使用占比呈逐年上升态势；SPSS、STATA 和 EVIEWS 这 3 类统计分析软件在功能上较为相似，其中 SPSS 和 EVIEWS 存在下降的趋势，而 STATA 则处于上升期，究其原因，SPSS 因为非常容易操作，初学者上手速度快，所以目前在 3 款相似软件

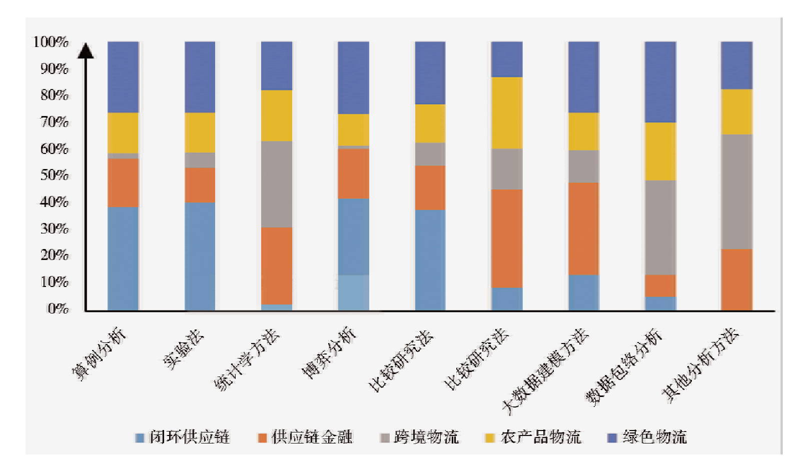


图 12 数据分析方法在不同研究热点的占比

表 10 使用次数前 20 名的数据分析工具

序号	名称	使用次数	序号	名称	使用次数
1	MATLAB	617	11	UCINET	27
2	SPSS	261	12	JAVA	20
3	AMOS	106	13	LISREL	16
4	STATA	92	14	SMARTPLS	15
5	ARCGIS	63	15	Python	15
6	EVIEWS	62	16	MAXDEA	14
7	LINGO	58	17	Frontier	14
8	CPLEX	54	18	GeoDa	12
9	VENSIM	45	19	VisualStudio	10
10	DEAP	43	20	CiteSpace	6

表 11 6 种软件在各年份中的使用情况

年份	MATLAB	SPSS	STATA	EVIEWS	ArcGIS	Python
2015 年	13.2%	7.3%	1.5%	1.8%	0.7%	0.0%
2016 年	12.8%	6.7%	1.1%	2.1%	0.7%	0.0%
2017 年	15.3%	6.3%	1.5%	1.5%	1.8%	0.1%
2018 年	16.7%	5.3%	3.0%	1.1%	2.2%	0.5%
2019 年	17.6%	5.2%	3.9%	0.9%	2.6%	1.5%

注:比例 = 当年使用该软件的论文/当年发表的全部论文

中使用率最高,但是在处理前沿的统计过程和数据管理范围上存在一些局限性,在学术节奏越来越快的当下,可能是导致其使用率在不断降低的原因之一,而 STATA 的操作也具有简单易懂的特点,同时在数据处理功能和数据管理能力上,随着该软件近几年的不断更新,也变得越发强大,因此该软件的使用率在不断追赶 SPSS,并隐隐有将其超过的势头。EVIEWS 虽然也是一款简单易上手的软件,但是扩展性较差,在需要大量编程的分析中,存在后劲不足的弱点,导致本身使用率不高的 EVIEWS 近几年还在不断下降。ArcGIS 和 Python 是近年来在物流领域的研究中开始使用的软件,使用率在不断上升。

值得一提的是,物流研究中使用数据分析工具多来自于理工类学科,例如使用 MATLAB 最多的学科是数学、计算机、电子等,SPSS、STATA、EVIEWS 的使用在计量经济学的研究中比较常见,ArcGIS 则在地理环境、旅游资源等研究中使用最多,Python 最先用于计算机学科的研究中,可以体现出物流研究乃至社会科学研究具有较强的学科交叉性。

5.2.3 模糊词使用分析

(1)模糊词使用情况概述。模糊词是指在描述科学数据时所使用的具有模糊概念的词语。模糊词的使用是否是因为研究中缺乏科学数据的支持所导致的?为了探究这个问题,本研究统计 8 种模糊词的使用情况,从表 12 可见,使用模糊词的论文占比 70%,说明模糊词的使用在物流领域的研究中比较常见。其中"大多""很多"在论文中出现的频率最高。为了进一步分析模糊词在每篇论文中的使用情况,统计 8 种模糊词的总使用次数和篇均使用次数见图 13。

表 12 模糊词类别及在样本中的使用情况

	论文数量/篇	类别	论文数量/篇	比例/%
模糊词	2 892	大量	1 793	62
		很多	1 299	45
		很少	568	20
		若干	459	16
		绝大多数	223	8
		少量	221	8
		差不多	27	1
		无数	23	1

从图 13 发现,"大量""很多"两词的总使用次数远超其他模糊词,在物流领域的研究中使用范围最广。"很少"的总使用次数要高于"若干",但篇均使用次数

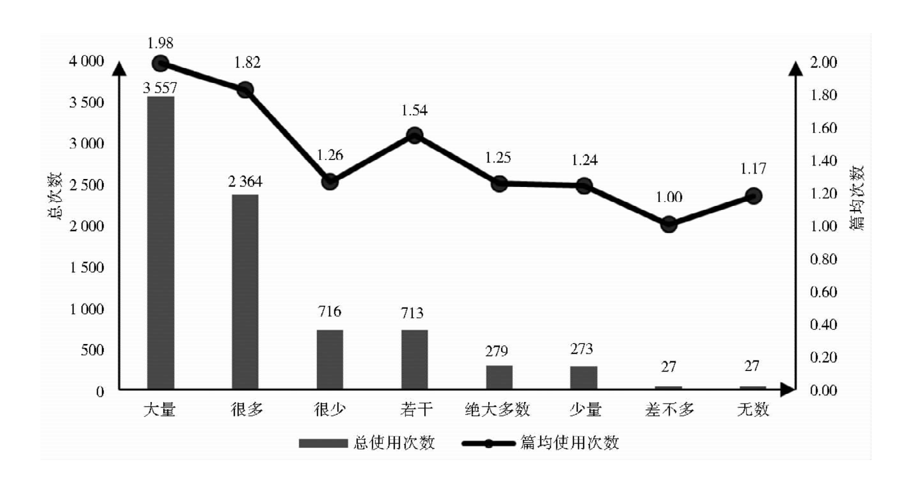


图 13　模糊词总使用次数及其篇均使用次数

却低于后者，说明“若干”的使用范围比较集中。而“差不多”和“无数”两词的受欢迎程度远没有其他 6 个词高，虽然都属于模糊词，但“差不多”这个词应该是模糊词中含义最“模糊”的一个了，“无数”本身绝对含义太深，这也是这两个词在学术环境越发严谨的当代使用率普遍较低的原因。

（2）不同年份下模糊词的使用。计算每一年使用模糊词的论文数量占全部论文的比例，见图 14，可以发现 8 种模糊词的使用次数在近 5 年呈缓慢下降趋势。

6　结论和启示

本研究对社会科学研究中物流研究领域进行分析后，有如下发现：

（1）在研究方法上，在前期研究的基础上，创新性提出了从样本文献中采集社会科学数据创建和使用相关信息的方法，包括在数据生命周期框架下利用迭代式方法构建“泛化－精确关键词词库”，能够保证数据采集的准确性与全面性，基于 Python 的二次匹配规则能够有效提高数据采集效率。通过对物流领域 4 114 篇 CSSCI 收录的文献进行社会科学数据相关信息的采集，并与前期研究成果进行比对，证明了这种方法的可行性和高效性。同时通过对数据的具体分析不难发现，物流研究在社会科学研究中有跨学科特性，研究对象的多样性也导致了其科学数据的复杂性，一定程度上反映了社会科学研究的复杂性。本方法在该领域的适用，为研究社会科学领域中其他学科的科学数据创建和使用提供了借鉴，为后续研究整个社会科学数据的创建和使用提供了可能，能够推动对社会科学数据的理解和认识，为制定更加合理有效的政策管理办法奠定基础。

（2）物流研究作为社会科学研究的一个研究领域，研究者在研究方法的使用和研究过程中有明显的跨学科特性；物流行业活动涉及到社会经济生活的各行各业，物流研究的对象并不局限于物流本身，而更加注重物流在人类社会经济活动中的结合，社会经济、产业链、供应链、组织、个体都是物流的研究对象，涉及到社会科学研究对象的大部分内容。因此，以物流科学数据作为社会科学数据的研究样本，具有较强的代表性，物流科学数据在一定程度上展现出了社会科学数据种类多、体量小、非标准化等基本特征。

（3）在数据来源的选择方面，按使用次数由高到低分别为科学调查数据、政府公开数据、商业公开数据，它们分别由科研机构与研究者、政府机构、行业协会与企业所创建。从使用次数上能够体现，科学调查数据因为能够通过问卷、访谈、调研等方式获取到更具研究针对性的数据，在研究中更受欢迎，但是其组织格式、存储方式更加复杂，获取难度也较大；除此以外，社会科学研究对政府公开数据重视度也较高，相比于商业公开数据获取更加容易、更具权威性，在使用次数上前者要远多于后者。在分析不同高产院校对数据来源的使用偏好时，发现地处北京的高校对政府公开数据的使用更多，说明位于首都的高校更加重视政府公开数据的利用，对政策有着更高的敏感度。

（4）社会科学数据的类型繁多，包括仿真数据、问卷数据等一手数据和政府文件数据、统计年鉴数据等二手数据，其中二手数据的使用要多于一手数据，这与近年来数据搜集方法的改变也有着密切关系，对数据搜集方法的分析，表明随着信息技术的不断推进，使用

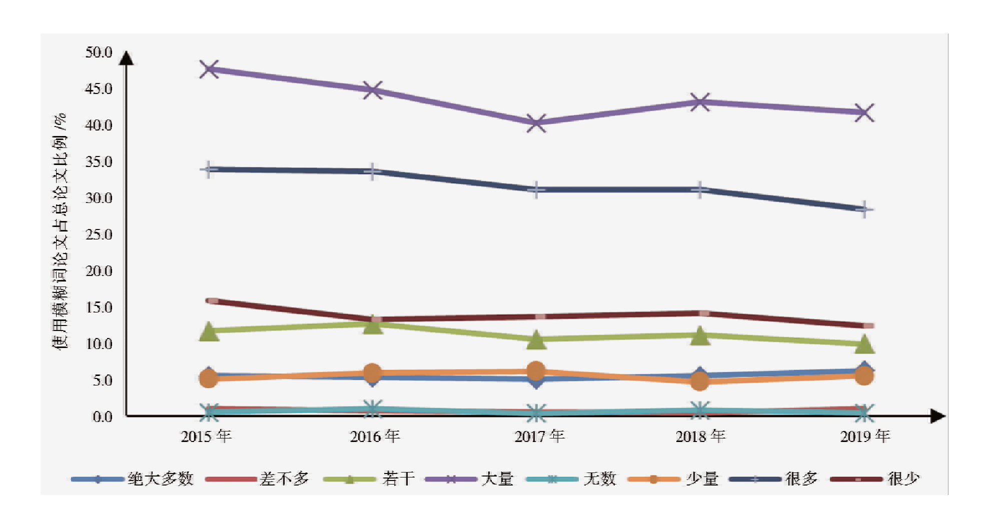


图 14 不同年份模糊词使用情况

网络查找搜集数据的论文在不断增加。值得一提的是，社会科学研究有较强学科交叉性，例如在物流的研究中经常会有多种不同类型的科学数据混用；数据搜集和分析工作在研究中变得越来越重要，为了提高研究效率，学者们倾向于 2 – 3 人合作完成研究工作。

(5)分析不同研究热点的数据分析偏好发现，不同研究主题对科学数据的偏好不同，这与其在数据分析方法的选择上有关，跨境物流、供应链金融、农产品物流、绿色物流更重视统计学方法，闭环供应链更偏好于算例分析、实验法和博弈分析。因此在后期的科学数据组织与管理上，可以依据研究领域、研究主题对社会科学数据进行分类管理。

(6)社会科学数据的分析工具繁多，但整体使用率不高，只有约 30% 的研究会使用，值得注意的是，研究中所使用的数据分析工具基本都是由国外研究机构或者学者所开发，说明我国的学术研究在数据分析工具的使用上存在较高的国外依赖性，需要加强自主研发，减少国外学术垄断的风险。

(7)通过模糊词分析，只有“大量”“很多”“很少”“若干”4 个模糊词的使用次数较多，由于模糊词本身具有不精确描述的特点，因此一个学科领域的研究使用如“大量”“很多”“很少”等模糊词，很容易被认为该学科的研究缺乏精确数据的支持，才导致大量模糊词的使用，但是从统计结果来看，无论是使用次数多还是少，8 种模糊词每一年的使用率都较为平稳，没有因为当下大数据、信息化时代的冲击而出现太大的波动，说明部分模糊词的频繁使用并不能表示一个学科领域的数据氛围不好，而很有可能只是与我国语言文化氛围和学者个人用语习惯有关[12]。

(8)文献计量分析仅作为外部环境分析方法，结合本文所提出的二次匹配方法采集论文中社会科学数据创建和使用的相关信息，在分析其特点的同时，从“内”和“外”两个角度对社会科学数据创建和使用进行深入研究。

本研究在前期研究基础上解决了从论文中提取社会科学数据创建和使用相关信息困难的问题，并利用物流研究领域文献进行分析。在后续研究中，一方面可以提高词库的完整性，在中文版基础上构建英文版词库，并依据不同学科进行调整；另一方面利用该方法对社会科学研究领域不同学科的研究论文开展大范围研究，同时将研究样本扩展到国际英文期刊，对不同学科、不同研究范式、不同研究领域间社会科学数据的创建和使用特点进行横向比较，全面展现我国社会科学数据的特征。

参考文献：

[1] PETERS I, KRAKER P, LEX E, et al. Zenodo in the spotlight of traditional and new metrics[J]. Frontiers in research metrics and analytics, 2017,2(13):1 – 14.
[2] HE L, NAHAR V. Reuse of scientific data in academic publications[J]. Aslib journal of information management, 2016, 68(4):478 – 494.
[3] 国务院办公厅. 国务院办公厅关于印发科学数据管理办法的通知[EB/OL]. [2020 – 03 – 20]. http://www.gov.cn/zhengce/content/201804/02/content_5279272.htm.
[4] 孙建军. 大数据时代人文社会科学如何发展[N]. 光明日报, 2014 – 07 – 07(11).
[5] 夏义堃. 人文社会科学数据管理的现实困境与对策分析[J]. 情报科学,2020, 38(9):14 – 22.

[6] BOLIKOWSKI L, HOUSSOS N, MANGHI P, et al. Data as "first-class citizens"[EB/OL]. [2020－08－13]. http://www.dlib.org/dlib/january15/01guest_editorial.html.
[7] NASA. Data & information policy[EB/OL]. [2021－01－23]. http://science. nasa. gov/earth-science/earth-science-data/data-information-policy/.
[8] BBSRC. BBSRC data sharing policy[EB/OL]. [2021－01－23]. http://www.bbsrc.ac.uk/about/policies/policy-foi/policy/data-sharing-policy/.
[9] 李志芳，邓仲华. 国内开放科学数据的分布及其特点分析[J]. 情报科学, 2015, 33(3): 45－49.
[10] 罗鹏程，崔海媛，赵静茹. 基于 DataCite 的科学数据现状特征研究[J]. 图书情报知识, 2019(3):101－112, 80.
[11] 孟祥保，钱鹏. 数据生命周期视角下人文社会科学数据特征研究[J]. 图书情报知识, 2017(1):76－88.
[12] 沈婷婷. 人文社科领域科学数据使用特征分析——基于《中国社会科学》样本论文的实证研究[J]. 大学图书馆学报, 2015, 33(3):101－107.
[13] MEADOWS A. To share or not to share? That is the (research data) question[EB/OL]. [2020－05－21]. http://scholarlykitchen.sspnet.org/2014/11/11/to-shareor-not-to-share-that-is-the-research-data-question.
[14] 谭春林，刘清海. 期刊编辑发表论文情况的文本挖掘与分析[J]. 编辑学报, 2019, 31(4): 407－410.
[15] 张娜，柳运昌，王若男. 基于文本情感分析的社交媒体数据挖掘[J]. 河南城建学院学报, 2019, 28(5):74－79.
[16] 刘玉林，菅利荣. 基于文本情感分析的电商在线评论数据挖掘[J]. 统计与信息论坛, 2018, 33(12):119－124.
[17] 任恒. 国内智库研究的知识图谱:现状、热点及趋势——基于 CSSCI 期刊(1998－2016)的文献计量分析[J]. 情报科学, 2018, 36(9):159－166.
[18] 冯亚飞，胡昌平，李霜双. 国内学术资源研究的知识图谱与热点主题[J]. 情报科学, 2019, 37(10):3－7, 19.
[19] 俞立平，王冰，张再杰. 历时扩散因子与历时相对扩散因子的应用研究——以 CSSCI 图书馆情报与文献学期刊为例[J]. 情报杂志, 2020, 39(3):156－162.
[20] 师荣华，刘细文. 基于数据生命周期的图书馆科学数据服务研究[J]. 图书情报工作, 2011,55 (1) :39－42.
[21] 丁宁，马浩琴. 国外高校科学数据生命周期管理模型比较研究及借鉴[J]. 图书情报工作, 2013, 57(6): 18－22.
[22] 武彤. 基于数据生命周期的美国研究图书馆科学数据开放共享服务研究[J]. 图书与情报, 2019 (1): 135－144.
[23] CEOS. Data life cycle models and concepts[EB/OL]. [2020－04－21]. http://www2.lib.virginia.edu/brown/data/.
[24] STARR J, WILLETT P, FEDERER P, et al. A collaborative framework for data management services: the experience of the university of California[EB/OL]. [2020－05－17]. https://escholarship.umassmed.edu/cgi/viewcontent.cgi? article=1014&context=jeslib.
[25] POUCHARD L. Revisiting the data lifecycle with big data curation[J]. International journal of digital curation, 2016, 10(2): 176－192.
[26] DCC. Curation lifecycle model[EB/OL]. [2020－07－12]. http://www.dcc.ac.uk/resources/curation-lifecycle-mode.
[27] UKDA. Research data lifecycle[EB/OL]. [2020－07－12]. http://www.data-archive.ac.uk/create-manage/life-cycle.
[28] 刘敏娟，张学福，颜蕴. 基于核心词、突变词与新生词的学科主题演化方法研究[J]. 情报杂志, 2016, 35(12):175－180.
[29] 肖明，陈嘉勇，李国俊. 基于 CiteSpace 研究科学知识图谱的可视化分析[J]. 图书情报工作, 2011, 55(6):91－95.
[30] 侯剑华，胡志刚. CiteSpace 软件应用研究的回顾与展望[J]. 现代情报, 2013, 33(4): 99－103.
[31] 陈悦，陈超美，刘则渊，等. CiteSpace 知识图谱的方法论功能[J]. 科学学研究, 2015, 33(2):242－253.
[32] 王发明，朱美娟. 国内区块链研究热点的文献计量分析[J]. 情报杂志, 2017, 36(12): 69－74, 28.
[33] 陈悦，陈超美，胡志刚，等. 引文空间分析原理与应用 CiteSpace 实用指南[M]. 北京:科学出版社, 2014.
[34] HOX J J, BOEIJE H R. Data collection, primary vs. secondary[J]. Encyclopedia of social measurement, 2005, 1 :593－599.
[35] 李华杰,史丹,马丽梅. 基于大数据方法的经济研究:前沿进展与研究综述[J]. 经济学家, 2018(6):96－104.
[36] 章成志，张颖怡. 基于学术论文全文的研究方法实体自动识别研究[J]. 情报学报, 2020, 39(6):589－600.
[37] 王芳，王向女. 我国情报学研究方法的计量分析:以 1999～2008 年《情报学报》为例[J]. 情报学报, 2010, 29(4):652－662.

作者贡献说明:

陈欣:对论文的选题、思路、撰写与修改进行指导和提出重要建议;

曹朝金:负责规则设计、程序编写、初稿撰写和论文修改;

叶春森:提供论文内容方向性修改意见;

汪传雷:对论文框架提出修改建议。

Research on Social Science Data Creation and Using
——Application of Twice Matching Data Acquisition Rules

Chen Xin[1] Cao Chaojin[2] Ye Chunsen[1] Wang Chuanlei[1]
[1] School of Business, Anhui University, Hefei 230009
[2] School of Management, Hefei University of Technology, Hefei 230009

Abstract: [Purpose/significance] Under the framework of the data life cycle, this paper proposes an innovative method for collecting information on the creation and use of social science data from papers, and deeply studies its basic situation, which provides a new idea for the research of social science data. [Method/process] Based on the papers collected by CSSCI from 2015 to 2020 in the field of logistics research with strong interdisciplinary intersection, this paper constructed thesauruses with generalized and accurate keyword based on the data life cycle through iterative method, collected the relevant information of social science data. Then, combined with the external environmental information of social science data, a comprehensive study of the creation and use of social science data has been carried out. [Result/conclusion] The rules is feasible and efficient in collecting information on the creation and use of social science data. Using Internet has become the main data collection method in social science research. Different research topics have different preferences for data use, and the popularity of data analysis tools is still low.

Keywords: social scientific data generalized-accurate thesaurus twice matching data acquisition rules Python bibliometrics